RoBoSTAR: Next-Scale Autoregressive Sign Language Translation for Humanoid Robots

Anonymous authors
Paper under double-blind review

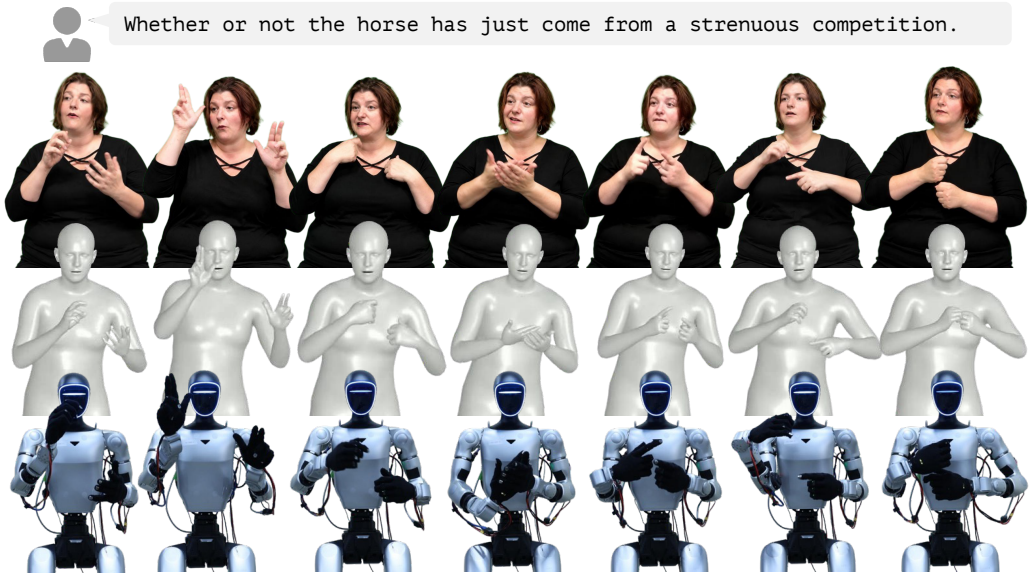


Figure 1: **From language to embodied robotic signing.** Given either speech or text, RoBoSTAR generates a continuous sign sequence and retargets it to a humanoid robot for physical execution.

Abstract

Sign-language interpretation in public communication relies on qualified professional interpreters and can be difficult to scale, motivating robotic signing as a complementary accessibility interface. We present RoBoSTAR, a text-conditioned sign language production (SLP) framework for generating human-centric sign motion that can be retargeted for robotic execution, with speech supported optionally through an external ASR front end. Conventional autoregressive approaches flatten motion into a single full-resolution token sequence, forcing long-range and local dependencies to be modeled at a uniform temporal granularity. RoBoSTAR instead combines part-wise Finite Scalar Quantization with next-scale autoregression, generating motion over progressively finer temporal resolutions while predicting synchronized body and hand tokens in parallel within each step. This coarse-to-fine formulation provides compact long-range context before progressively refining motion details, while self-conditioning and context corruption improve robustness to cross-scale prediction errors. The generated motion is subsequently retargeted for physical humanoid execution. Extensive qualitative and quantitative evaluations are conducted to demonstrate the effectiveness of RoBoSTAR.

1 Introduction

Sign languages are natural languages used by many deaf and hard-of-hearing (DHH) people, conveying meaning through coordinated handshapes, body motion, and non-manual markers such as facial

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

expressions and mouth movements. In public broadcasts, emergency briefings, and press conferences, qualified professional interpreters remain essential. Robotic signing is therefore not intended to replace professional interpretation, but may serve as a complementary accessibility channel for selected applications and resource-constrained settings.

Recent work has advanced sign language production (SLP) from continuous motion regression to discrete language-model-based generation, while robotic systems have begun to explore embodied sign interaction (Qiao et al., 2026; Tan et al., 2026). We develop a text-conditioned sign language motion generator that produces human-centric body and bilateral-hand motion. The generated motion can then be transferred to a humanoid robot through morphology-specific retargeting and control, providing a practical path from SLP to physical execution. Speech is supported optionally through a frozen off-the-shelf ASR (Automatic Speech Recognition) front end. Although the human-centric generation includes facial and other non-manual cues, the target robot lacks facial and lip actuation; these cues are therefore omitted during embodiment, limiting the intelligibility of signs that depend on non-manual markers. As illustrated in Fig. 1, RoBoSTAR translates sign motion from language and retargets it to a humanoid robot for physical execution.

Existing sign language production (SLP) methods span both continuous and discrete motion generation paradigms. Continuous regression and diffusion approaches directly model continuous motion trajectories, while diffusion-based methods generate motion through iterative denoising over multiple sampling steps (Saunders et al., 2020b; Hwang et al., 2021; Baltatzis et al., 2024). Recent discrete approaches instead tokenize motion using VQ-based representations (van den Oord et al., 2017) and autoregress over flat temporal sequences (Yin et al., 2024a; Zuo et al., 2025; Symeonidis-Herzig et al., 2026). Such representations provide a natural interface for pretrained language models (Jiang et al., 2023; Zuo et al., 2025), but learned VQ codebooks can exhibit uneven utilization, while flat next-token decoding entangles long-range and local dependencies within a single full-resolution autoregressive chain, imposing a uniform temporal granularity throughout generation.

To address these limitations, we present RoBoSTAR, a hierarchical autoregressive framework for embodied SLP. RoBoSTAR constructs synchronized body and bilateral-hand token streams using part-wise FSQ (Mentzer et al., 2024). Unlike VQ methods that perform nearest-neighbor lookup in a learned codebook, FSQ quantizes bounded scalar values over a fixed grid, providing a regular discrete space for language modeling. On this representation, an LM backbone (Xue et al., 2021) performs *next-scale autoregression*, inspired by coarse-to-fine generation (Tian et al., 2024; Zheng et al., 2026; Liu et al., 2025). Coarser temporal scales provide compact long-range context, while subsequent scales progressively refine the motion at higher temporal resolution. Within each scale, temporal positions remain autoregressive, while synchronized body and bilateral-hand tokens are predicted in parallel.

Since errors in the coarse scale can be propagated through subsequent refinement stages, we introduce self-conditioning (SC) and generic context corruption (GCC). Both strategies expose the model to imperfect preceding-scale contexts during training, reducing the train–test mismatch and improving robustness to errors propagated from coarser predictions.

Existing geometric metrics primarily measure spatial agreement with the reference and can therefore under-characterize temporal degeneration, such as near-static predictions or frozen terminal segments. We introduce three complementary dynamics metrics—Dynamic Degree, Magnitude Ratio, and Freeze-tail—to explicitly measure motion activity, motion magnitude, and terminal freezing.

The generated human-centric motion is transferred to the target robot through morphology-specific retargeting and tracking (Qiao et al., 2026; Chang et al., 2024), keeping sign generation independent of robot morphology.

Overall, our contributions are as follows:

- We propose a part-wise FSQ next-scale autoregressive model for sign language motion generation, jointly modeling body and bilateral-hand motion across progressively finer temporal resolutions.
- We introduce self-conditioning (SC) and generic context corruption (GCC) to mitigate cross-scale exposure bias and improve robustness.
- We introduce Dynamic Degree, Magnitude Ratio, and Freeze-tail to characterize degenerate motion dynamics that are not captured by conventional geometric or semantic metrics.

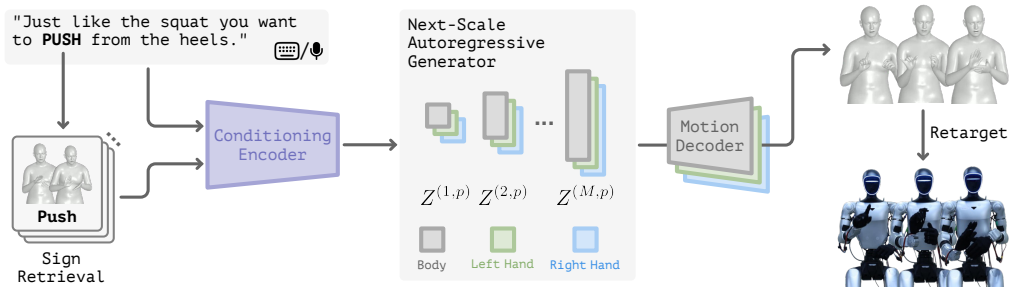


Figure 2: **Overview of RoBoSTAR.** Speech is first transcribed into text, while written input directly conditions the generator with retrieved signs. Part-wise FSQ tokenization represents sign motion at multiple temporal resolutions, which are generated sequentially through next-scale autoregression. The decoded human motion is subsequently retargeted to the robot in the real world.

2 Related Works

Sign Language Generation. Sign language generation (SLG) maps natural-language inputs to human signing motion. Early methods generate continuous pose sequences from text or gloss (Saunders et al., 2020b;a; Hwang et al., 2021; Ma et al., 2024; Stoll et al., 2022; Gil-Martín et al., 2023), later extending from sparse keypoints to holistic 3D avatars in both sentence- and word-level settings (Baltatzis et al., 2024; Yu et al., 2024; Dong et al., 2024). Recent work discretizes motion for language-model-based generation: MotionGPT (Jiang et al., 2023) establishes this paradigm for general human motion, T2S-GPT (Yin et al., 2024a) applies autoregressive discrete modeling to SLG, SOKE (Zuo et al., 2025) combines part-wise body–hand tokenization with a pretrained multilingual LM and sign retrieval, and M^{3}T (Symeonidis-Herzig et al., 2026) further models body, hand, and facial motion with modality-specific tokens. We exclude M^{3}T from quantitative comparison because our robot lacks facial actuation and the code is not publicly released.

Complementary work targets visual synthesis: SignLLM (Fang et al., 2025a) generates pose sequences from text or prompts for subsequent video rendering, while other methods perform pose-conditioned (Stoll et al., 2020; Fang et al., 2025b) or end-to-end sign-video generation (Qi et al., 2024; Fang et al., 2026). These works primarily target human sign visualization, making them less suitable for scenarios that require sign language to be physically performed on-site.

Autoregressive and Next-Scale Motion Generation. Conventional next-token autoregression models motion through a single full-resolution temporal sequence, entangling long-range and local dependencies under a uniform temporal granularity. Next-scale autoregression instead generates global structure before local detail. VAR (Tian et al., 2024) introduces coarse-to-fine scale prediction for images; Infinity (Han et al., 2025) scales this paradigm through bitwise prediction, and InfinityStar (Liu et al., 2025) extends it to unified spatio-temporal image and video generation. For human motion, MoScale (Zheng et al., 2026) predicts progressively finer temporal resolutions so that global semantics are established before local dynamics. RoBoSTAR brings this principle to SLG by jointly modeling body and hand streams across progressively finer temporal resolutions, using coarser tokens for compact long-range context and finer tokens for higher-resolution refinement.

Embodied Robotic Signing. Existing robotic signing systems address related but distinct tasks. Human-to-robot tactile ASL transfer (Chang et al., 2024) retargets human demonstrations to a bimanual robot, while SignVLA (Tan et al., 2026) maps finger-spelled signs to manipulation commands. SignBot (Qiao et al., 2026) focuses on bidirectional human–robot sign interaction and largely integrates established components for sign understanding, response generation, sign generation, retargeting, and control. Its sign generator follows SOKE (Zuo et al., 2025). In contrast, we formulate robotic signing as an embodied translation task and focus on advancing sign generation itself, introducing part-wise FSQ, next-scale autoregression, and robust cross-scale conditioning before robotic retargeting and execution.

3 METHOD

We present **ROBOSTAR** for robotic signing, as illustrated in Fig. 2. The input can be either speech or text; speech is first transcribed into text by a frozen, off-the-shelf ASR front end. ROBOSTAR then translates the text into human-centric sign motion through next-scale autoregressive generation. The final output is the corresponding humanoid robot motion for sign-language translation.

3.1 PART-WISE FSQ MOTION REPRESENTATION

We represent each normalized human-centric sign sequence as $S = [S^B, S^{LH}, S^{RH}] \in \mathbb{R}^{T\times d}$, $d = d_B + d_{LH} + d_{RH}$, where T denotes the sequence length and $S^p \in \mathbb{R}^{T\times d_p}$ is the motion stream for $p \in \mathcal{P} = \{B, LH, RH\}$, corresponding to the body, left-hand, and right-hand streams, respectively. The body stream additionally includes facial motion parameters. Because body and hand motions exhibit distinct statistics, a shared tokenizer may underrepresent subtle manual articulation. We therefore employ dedicated temporal encoders and decoders for each part, as illustrated in Fig. 3.

Given S^p, the corresponding encoder produces the finest-resolution continuous latent sequence

$$H^p = \mathcal{E}_p(S^p) \in \mathbb{R}^{L_M\times C}. \tag{1}$$

Each latent vector is projected into a low-dimensional FSQ space, $u^p_\tau = \mathcal{P}^{\text{in}}_p(h^p_\tau) \in \mathbb{R}^{d_q}$, and quantized independently along each scalar dimension: $\hat{u}^p_{\tau,j} = Q_{n_{p,j}}\left(u^p_{\tau,j}\right), \qquad K_p = \prod_{j=1}^{d_q} n_{p,j}$. The quantized tuple is converted through mixed-radix encoding into a finest-scale token $z^{(M,p)}_\tau \in \{1, \ldots, K_p\}$. Gradients through the rounding are estimated using a straight-through estimator.

Coarser token sequences are constructed directly from the finest-scale tokens, without decoding and re-encoding motion. We first dequantize the finest-scale IDs into their corresponding C-dimensional implicit FSQ code vectors:

$$E^{(M,p)} = \text{Dequantize}_p\left(Z^{(M,p)}\right) \in \mathbb{R}^{L_M\times C}. \tag{2}$$

For temporal divisors $(d_1, \ldots, d_M)$, the target length of scale m is $L_m = \max\left(1, \left\lceil \frac{L_M}{d_m} \right\rceil\right)$. For each scale $m < M$, the finest-scale code vectors are resized directly to L_m using one-dimensional adaptive average pooling and projected back onto the fixed set of valid FSQ codes:

$$\bar{E}^{(m,p)} = \text{AdaptiveAvgPool}\left(E^{(M,p)}, L_m\right), \quad z^{(m,p)}_\tau = \arg\min_{k\in\{1,\ldots,K_p\}} \left\| \bar{e}^{(m,p)}_\tau - c^p_k \right\|^2_2. \tag{3}$$

where $\{c^p_k\}_{k=1}^{K_p}$ denotes the fixed set of dequantized FSQ codes for part p. Adaptive pooling produces exactly L_m output positions for arbitrary L_M, without truncation or zero padding. The resulting hierarchy is $Z = \{Z^{(m,p)}\}_{m=1,\ldots,M;\, p\in\mathcal{P}}$, with $L_1 < \cdots < L_M$. All coarser scales are independently derived from $Z^{(M,p)}$; the construction is therefore neither recursive nor residual. Exact pooling boundaries for non-divisible sequence lengths are given in Appendix A.1.

Unlike VQ-based tokenizers, finest-scale FSQ encoding does not require nearest-neighbor lookup in a learned embedding codebook. Its fixed Cartesian grid avoids learned-codebook utilization issues and provides a regular prediction space. The nearest-code operation above is used only to project pooled vectors back onto the valid FSQ grid when constructing coarse-scale targets. We adopt a compact quantization space to balance reconstruction fidelity and downstream token predictability.

The complete hierarchy supervises coarse-to-fine generation, whereas only the finest-scale tokens are decoded into motion. For each part, $\widetilde{H}^p = \text{Dequantize}_p\left(Z^{(M,p)}\right), \quad \hat{S}^p = \mathcal{D}_p(\widetilde{H}^p)$, and the reconstructed streams are concatenated as $\hat{S} = [\hat{S}^B, \hat{S}^{LH}, \hat{S}^{RH}]$.

3.2 NEXT-SCALE SIGN LANGUAGE GENERATION

Let x denote either a written input or the transcript produced by the frozen ASR front end. ROBOSTAR employs a pretrained multilingual language-model backbone to extract text features. Following retrieval-enhanced SLP (Zuo et al., 2025), we augment the language condition with dictionary-based lexical sign retrieval. Input words are matched against the dictionary in sentence order, and at most three unique matches are retained.

Their body and bilateral-hand FSQ prototypes are serialized as lexical motion cues into the conditioning prompt, yielding C_x. The complete sentence-level motion is still generated autoregressively rather than assembled from retrieved sign clips. Further details on dictionary construction, lexical matching, and retrieval augmentation are provided in Appendix A.2. A length predictor estimates the finest-scale token length L_M, while coarser-scale lengths are determined using fixed temporal ratios.

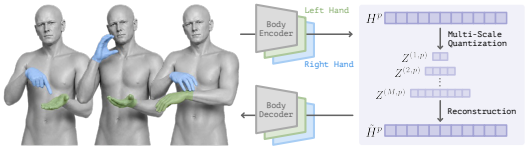


Figure 3: **Part-wise FSQ pyramid.** Finest-scale tokens are obtained from each motion stream, while coarser targets are constructed by temporal pooling and projection onto the FSQ grid. The hierarchy supervises next-scale generation, and only the finest scale is decoded for reconstruction.

Rather than flattening all motion tokens into a single sequence, RoBoSTAR factorizes generation causally across M temporal scales:

$$p_\theta(Z \mid x) = \prod_{m=1}^{M} \prod_{\tau=1}^{L_m} \prod_{p \in \mathcal{P}} p_{\theta,p}\left(z_\tau^{(m,p)} \mid C_x,\, Z^{(<m)},\, Z_{<\tau}^{(m)}\right), \tag{4}$$

where $Z^{(<m)} = \{Z^{(j,p)} \mid j < m,\ p \in \mathcal{P}\}$ contains all previously generated coarser scales, and $Z_{<\tau}^{(m)} = \{z_{\tau'}^{(m,p)} \mid \tau' < \tau,\ p \in \mathcal{P}\}$ denotes the prefix of the current scale. Here, next-scale autoregression refers to causal dependency across temporal resolutions: scales are generated sequentially from coarse to fine, while temporal positions remain autoregressive within each scale. At each position, the shared decoder hidden state is projected onto disjoint body, left-hand, and right-hand motion vocabularies, producing the three synchronized tokens in parallel.

For scale $m > 1$, let $E_k \in \mathbb{R}^{L_k \times D}$ denote the motion-token embedding sequence obtained from a preceding scale $k < m$. We align it to the current length L_m by deterministic zero-order index resampling: $[\mathcal{A}_{k\to m}(E_k)]_t = (E_k)_{\lfloor tL_k/L_m \rfloor}, \qquad t = 0, \ldots, L_m - 1$. Thus, coarse-scale embeddings are repeated over contiguous temporal regions rather than linearly interpolated.

For each ordered scale pair (m, k) with $k < m$, we learn a scalar gate $g_{m,k} = \sigma(a_{m,k})$, where $a_{m,k}$ is a learned gate logit. The aligned preceding-scale representations are fused into the cross-scale context

$$H_m = W_m \left(\frac{\sum_{k<m} g_{m,k}\, \mathcal{A}_{k\to m}(E_k)}{\sum_{k<m} g_{m,k}} \right), \tag{5}$$

where W_m is a learned scale-specific linear projection. The resulting context is added to the current-scale decoder representations. Standard causal self-attention restricts each position to the current-scale prefix, while the language condition is accessed through encoder–decoder cross-attention. Further implementation details are provided in Appendix A.1.

Standard teacher forcing provides clean preceding-scale contexts during training, whereas inference conditions finer scales on previously predicted scales. To reduce this cross-scale exposure bias, we apply self-conditioning (SC) followed by generic context corruption (GCC) to preceding-scale contexts. For SC, each completed scale is replaced sample-wise by its teacher-forced argmax prediction with probability p_{SC}; the replacement jointly covers body and bilateral-hand tokens. GCC then independently replaces each valid context token with a uniformly sampled code from the corresponding part vocabulary with probability p_{GCC}. Both probabilities follow linear warm-up schedules during training; full specifications are provided in Appendix A.3.

The generator is optimized with scale- and part-weighted token prediction losses, together with auxiliary length prediction and text–motion contrastive objectives:

$$\mathcal{L} = \sum_{m=1}^{M} \lambda_m \sum_{p \in \mathcal{P}} \alpha_p \mathcal{L}_{\mathrm{CE}}^{(m,p)} + \lambda_{\mathrm{len}} \mathcal{L}_{\mathrm{len}} + \lambda_{\mathrm{con}} \mathcal{L}_{\mathrm{con}}. \tag{6}$$

Here, $\mathcal{L}_{\mathrm{len}}$ supervises the finest-scale motion length, while $\mathcal{L}_{\mathrm{con}}$ aligns text and motion representations through a symmetric contrastive objective. Exact loss definitions, weighting coefficients, and training hyperparameters are provided in Appendix A.4. During inference, RoBoSTAR generates scales sequentially from coarse to fine, after which only the finest-scale tokens are decoded into continuous motion $\hat{S}$.

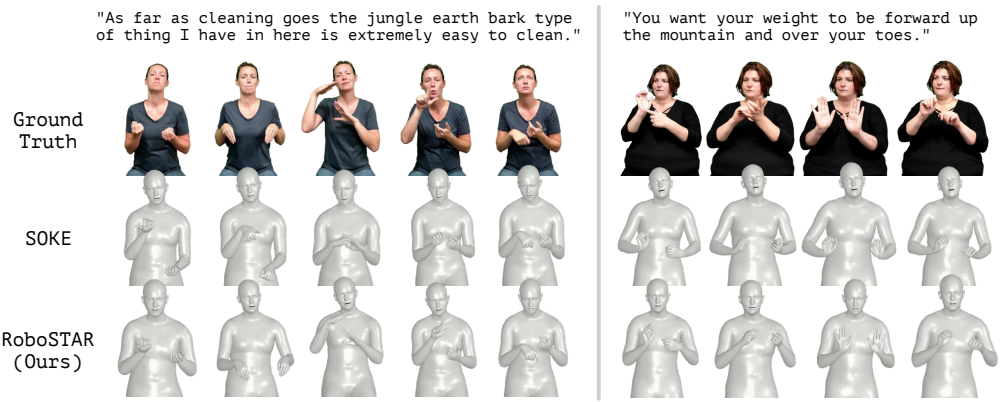


Figure 4: **Qualitative comparison on How2Sign.** Representative frames from ground-truth (GT), SOKE, and RoBoSTAR (Ours) sequences. RoBoSTAR better preserves the reference upper-body motion, hand articulation, and body–hand coordination.

Table 1: Quantitative comparison on the common How2Sign test subset. GT is reported only as a reference and is excluded from ranking. B-n denotes back-translation BLEU-n. Best and second-best results among the generation methods are shown in bold and underlined.

Method	DTW-PA-JPE ↓		DTW-JPE ↓		B-T BLEU ↑				Motion Dynamics		
	Body	Hand	Body	Hand	B-1	B-2	B-3	B-4	Dynamic ↔ 0.8	Magnitude ↔ 1	Freeze-tail ↓
GT	–	–	–	–	22.45	8.59	3.95	1.96	0.804	1.000	0.036
PT	10.85	2.66	10.84	10.97	13.44	4.69	1.99	0.91	0.245	0.484	0.232
NAR	4.21	1.79	4.14	5.70	6.81	2.10	0.88	0.42	0.000	0.004	1.000
S-MotionGPT	11.84	3.92	11.91	13.30	12.01	4.33	1.99	1.03	0.005	0.016	0.294
T2S-GPT	4.35	1.82	4.27	5.78	7.13	2.79	1.28	0.60	0.000	0.000	1.000
SOKE	5.73	2.08	5.73	7.26	17.25	6.31	2.86	**1.39**	0.388	0.515	**0.163**
Ours	**3.96**	**1.52**	**3.87**	**5.34**	**20.08**	**7.12**	**3.04**	1.36	**0.432**	**0.523**	0.187

3.3 Embodied Retargeting and Tracking

RoBoSTAR generates human-centric sign motion rather than robot joint commands. From the decoded sequence, we recover SMPL-X upper-body trajectories (Pavlakos et al., 2019) and MANO hand articulation (Romero et al., 2017). Following task-space human-to-robot transfer (Qiao et al., 2026; Chang et al., 2024), the upper-body motion is mapped to the target morphology in a torso-centered frame under joint-limit, reachability, self-collision, and temporal-continuity constraints. Hand articulation is retargeted independently from palm-normalized MANO geometry, and the resulting body and hand references are tracked by a platform-specific low-level controller. Motion channels unsupported by the target embodiment, including facial expression parameters, are omitted during robotic execution. This modular interface keeps sign generation independent of robot morphology; additional details on robotic retargeting and quantitative embodiment evaluation are provided in Appendix A.5 and Appendix A.6, respectively.

4 Experiments

4.1 Experimental Setup

Data Preparation. We evaluate RoBoSTAR on How2Sign (Duarte et al., 2021), one of the largest publicly available datasets for continuous ASL and a widely used benchmark for sign language generation, containing approximately 35K video–text pairs. We use the official training, validation, and test splits. For benchmark evaluation, the paired English transcripts are used as input

to isolate motion generation from upstream speech-recognition errors. We reconstruct 3D motion using WHAC (Yin et al., 2024b) for the upper body and WiLoR (Potamias et al., 2025) for MANO-compatible hand articulation. Each sequence is represented as $S \in \mathbb{R}^{T \times d}$, where T is the number of frames and $d = 133$ denotes the per-frame dimensionality, comprising body and bilateral-hand pose parameters. We discard camera, lower-body, and identity-dependent shape that are irrelevant or unsupported by the target robot.

Evaluation Metrics. Following prior 3D sign language production work (Zuo et al., 2025; Symeonidis-Herzig et al., 2026), we evaluate geometric fidelity using DTW-JPE and DTW-PA-JPE on native-length sequences. DTW-JPE computes root- or wrist-aligned joint errors, whereas DTW-PA-JPE additionally applies per-frame similarity Procrustes alignment. We report cumulative DTW alignment cost in meters, averaging the left- and right-hand errors.

To complement reference-based geometric metrics, we additionally measure whether a generated sequence exhibits realistic temporal activity. Dynamic Degree measures the fraction of sufficiently active transitions, Magnitude Ratio compares the overall motion magnitude of generated and ground-truth sequences, and Freeze-tail detects prolonged low-motion segments at the end of a sequence. Dynamic Degree and Magnitude Ratio have GT-derived reference values of approximately 0.8 and 1.0, respectively, while lower Freeze-tail is preferred.

For semantic evaluation, we use a separately trained and frozen Hybrid FSQ-T5 back-translator. Since the evaluator is itself imperfect, its BLEU scores are interpreted comparatively rather than as an absolute measure of translation accuracy. Full metric definitions, units, and back-translation details are provided in Appendix A.7.

Implementation Details. We train separate part-wise FSQ tokenizers with $4\times$ temporal down-sampling and initialize the generator from mT5-Large (Xue et al., 2021), using four temporal divisors $(8, 4, 2, 1)$. Speech, when provided, is transcribed using the frozen Granite Speech 4.1 model (Team, 2026). Generated motions are retargeted to Unitree G1 with dual 20-DoF Wuji hands. Full optimization and training details are provided in Appendix A.4.

4.2 Comparison with State-of-the-Art Methods

Quantitative Results. We evaluate all available predictions on the common test subset using the same joint topology, temporal alignment, and metric implementation. As shown in Table 1, RoBoSTAR achieves the lowest errors across all four geometric metrics, outperforming NAR, the strongest prior baseline on geometric accuracy under our unified evaluation protocol, on both body and hand motion. Notably, NAR achieves competitive geometric errors while exhibiting near-zero Dynamic Degree and Magnitude Ratio together with a Freeze-tail of 1.0, revealing severely degenerate, near-static motion. This contrast shows that favorable geometric errors alone can mask motion collapse, motivating the complementary dynamics metrics to expose this failure mode. RoBoSTAR also substantially improves all geometric metrics over the closely related SOKE baseline.

For semantic fidelity, RoBoSTAR achieves the best B-1, B-2, and B-3 scores and the second-best B-4 score. It further produces the Dynamic Degree and Magnitude Ratio closest to their reference targets, while attaining the second-lowest Freeze-tail value. Together, these results indicate that RoBoSTAR improves geometric and semantic fidelity while maintaining substantially healthier motion dynamics.

Qualitative Results. Figure 4 compares representative sequences generated by RoBoSTAR and SOKE against the ground truth. RoBoSTAR more closely follows the reference upper-body trajectories and produces clearer hand articulation and body–hand coordination.

Real-Robot Execution. Figure 5 presents representative executions of RoBoSTAR-generated motions after morphology-aware retargeting. The physical robot reproduces coordinated upper-body and dexterous-hand movements, demonstrating the feasibility of transferring the generated human-centric motions to a robotic platform.

Physical Embodiment Evaluation. We further quantify the distortion introduced by robotic embodiment. Retargeting introduces measurable morphology-transfer error, while the generated motions remain close to GT in positional feasibility, with elbow and wrist errors of $2.0\,\text{cm}$ and $5.6\,\text{cm}$ for RoBoSTAR, compared with $2.0\,\text{cm}$ and $4.6\,\text{cm}$ for GT motion. Wrist orientation exhibits a

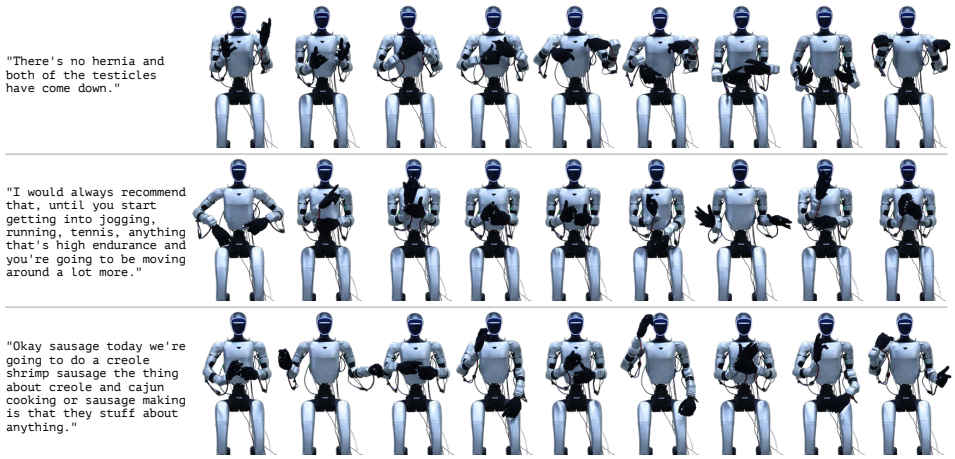


Figure 5: **Real-robot execution of ROBOSTAR-generated sign motion.** Representative frames show generated human-centric motions transferred to the physical robot through morphology-aware retargeting and low-level tracking.

Table 3: Reconstruction–generation trade-off between FSQ and VQ-VAE at a matched compression ratio, comparing tokenizer reconstruction fidelity and downstream motion generation.

Quantizer	Reconstruction		Generation			
	Norm. MAE	Norm. RMSE	DTW-PA Body	DTW-PA Hand	DTW Body	DTW Hand
FSQ	0.5221	0.7053	**3.9643**	**1.5238**	**3.8672**	**5.3448**
VQ-VAE	**0.4098**	**0.5569**	4.7488	2.1826	4.6157	6.3220

larger residual, while closed-loop physics simulation shows that the retargeted references remain trackable with a 2.1 cm wrist error and 1.3° joint-angle MAE. Overall, these results indicate that the generated motions remain reasonably compatible with the target robot morphology despite non-zero retargeting error. Full definitions and results are provided in Appendix A.6.

4.3 Ablation Studies

Effect of Next-Scale Generation. To isolate the contribution of next-scale autoregression, we compare the complete four-scale hierarchy $\{8, 4, 2, 1\}$ with a conventional single-scale autoregressive model operating only at full temporal resolution $\{1\}$. Both variants use the same backbone, tokenizer, and training protocol, differing only in the temporal generation hierarchy.

As shown in Table 2, next-scale generation consistently improves all geometric metrics. Compared with single-scale autoregression, it reduces DTW-PA-JPE by 14.9% for the body and 40.4% for the hands, and reduces DTW-JPE by 12.7% and 21.5%, respectively. These results demonstrate that coarse-to-fine temporal generation provides a more effective formulation than flat full-resolution autoregression, particularly for fine-grained hand motion.

Table 2: Effect of next-scale generation.

Variant	DTW-PA-JPE ↓		DTW-JPE ↓	
	Body	Hand	Body	Hand
Single-scale AR	4.66	2.56	4.43	6.81
Next-scale	**3.96**	**1.52**	**3.87**	**5.34**

Motion Tokenization. Table 3 compares FSQ and VQ-VAE under the same part-wise architecture and a matched compression ratio of $740.17\times$. VQ-VAE achieves lower oracle reconstruction error, whereas FSQ consistently yields better downstream generation across all geometric metrics. This result highlights a reconstruction–generation trade-off: the tokenizer with higher reconstruction fidelity does not necessarily provide the most predictable targets for autoregressive generation. Despite its slightly weaker reconstruction, the regular FSQ space leads to more accurate generated motions. Additional ablations on the language-model backbone and sign retrieval are provided in Appendix A.8.

Table 4: Ablation study of self-conditioning (SC) and generic context corruption (GCC). The SC+GCC configuration is adopted as our final model. Best and second-best results are shown in bold and underlined.

Configuration	Training		DTW-PA-JPE ↓		DTW-JPE ↓		B-T BLEU ↑				Motion Dynamics		
	SC	GCC	Body	Hand	Body	Hand	B-1	B-2	B-3	B-4	Dynamic ↔ 0.8	Magnitude ↔ 1	Freeze-tail ↓
Neither	✗	✗	4.19	1.85	4.06	5.90	8.85	3.02	1.37	0.63	0.13	0.23	**0.16**
GCC only	✗	✓	4.25	1.96	4.10	6.04	8.17	2.74	1.23	0.55	0.15	0.25	0.23
SC only	✓	✗	**3.92**	1.54	**3.82**	**5.07**	17.93	6.34	2.67	1.14	0.34	0.47	0.31
Ours (SC+GCC)	✓	✓	3.96	**1.52**	3.87	5.34	**20.08**	**7.12**	**3.04**	**1.36**	**0.43**	**0.52**	0.19

Self-conditioning and context corruption. Table 4 studies self-conditioning (SC) and generic context corruption (GCC). GCC alone degrades geometric accuracy, whereas SC substantially improves it by exposing finer scales to model-generated histories. Adding GCC to SC introduces a modest geometric trade-off but substantially improves back-translation and motion-dynamics metrics. We therefore adopt SC+GCC as the final configuration, balancing geometric fidelity, semantic consistency, and motion robustness.

4.4 User Study

We conduct a blinded user study with four fluent signers using 15 held-out sentences from the How2Sign dataset. For each sentence, participants view the ground-truth motion, SOKE output, and RoBoSTAR output in randomized order with method identities hidden. Each motion receives a single overall score on a 1–10 scale, jointly considering semantic accuracy, hand-shape clarity, motion naturalness, and temporal fluency. Ratings are first averaged across the four participants for each sentence, and we report the mean and standard error of the mean (SEM) across the resulting 15 sentence-level scores.

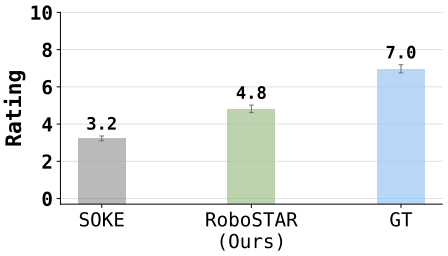


Figure 6: **User study.** Mean ratings; error bars denote SEM.

As shown in Fig. 6, RoBoSTAR receives higher mean ratings than SOKE and narrows the perceptual gap to the ground-truth motions. Together, these ratings provide complementary evidence that RoBoSTAR is a promising sign-generation backbone for downstream applications, including public information delivery and embodied accessibility systems.

5 Conclusion

In this paper, we introduced RoBoSTAR, a framework for robotic sign language translation. At its core, RoBoSTAR combines part-wise FSQ with next-scale autoregression to generate synchronized body and hand motion across progressively finer temporal resolutions, while self-conditioning and generic context corruption improve robustness to cross-scale prediction errors. We further introduce three complementary motion-dynamics metrics to characterize failure modes that are not captured by conventional geometric or semantic evaluation. Under a unified evaluation protocol on reconstructed 3D motion from the common How2Sign test subset, RoBoSTAR achieves state-of-the-art geometric performance. Physical robot demonstrations further illustrate the feasibility of executing the generated motions on a humanoid platform, promising future embodied accessibility applications for DHH communities.

Limitations and Future Work. Our system is currently evaluated only on How2Sign for ASL, leaving cross-domain and cross-language generalization for future study. Training motions are reconstructed from monocular videos, making fine-grained hand articulation and occlusions sensitive to 3D reconstruction errors. In addition, the current robot lacks facial and lip actuation for non-manual markers. Future work will broaden data and language coverage, improve 3D supervision, and explore richer robotic embodiments.

AI USE STATEMENT

In this work, generative AI tools were used in two limited capacities. First, AI-based enhancement was applied to improve the visual clarity of blurred source videos for visualization purposes only; the enhanced videos were not used to produce the reported quantitative results. Second, generative AI was used for language polishing and editorial assistance during manuscript preparation. All AI-assisted outputs were reviewed and, where necessary, revised by us. Generative AI was not used to generate experimental results, quantitative evaluations, human-study responses, or scientific conclusions. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work explores robotic signing as a complementary accessibility technology rather than a replacement for qualified professional sign language interpreters, particularly in emergency, medical, legal, or other high-stakes settings where mistranslation may have serious consequences. The current system is evaluated on ASL data from How2Sign and may inherit dataset- and signer-specific biases. Moreover, the target robot cannot reproduce facial and lip-based non-manual markers, which may reduce intelligibility for signs that depend on these cues. These limitations should be considered before any real-world deployment.

Our user study involved four fluent signers. All participants were informed about the study procedure and the intended research use of their responses, and voluntarily consented to participate. No participant-identifying information is reported.

We use the system only for research evaluation and do not claim that it is ready for unsupervised deployment in safety-critical communication.

REPRODUCIBILITY STATEMENT

We provide implementation and evaluation details throughout the main paper and appendix to facilitate reproducibility. The data processing, model configuration, and main experimental protocol are described in Sec. 4. Additional implementation details for multi-scale construction and cross-scale conditioning, sign retrieval, self-conditioning and context corruption, the training objectives, and robotic retargeting are provided in Appendix A.1, Appendix A.2, Appendix A.3, Appendix A.4, and Appendix A.5, respectively. Complete definitions and units for the geometric and motion-dynamics metrics, together with the architecture, data splits, and held-out performance of the back-translation evaluator, are provided in Appendix A.7.

REFERENCES

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text. In *Conference on Computer Vision and Pattern Recognition*, 2024.

Che-Ming Chang, Felipe Sanches, Geng Gao, and Minas Liarokapis. On human to robot skill transfer for the execution of complex tactile american sign language tasks with a bimanual robot platform. In *46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC*, 2024.

Lu Dong, Xiao Wang, and Ifeoma Nwogu. Word-conditioned 3d american sign language motion generation. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Amanda Cardoso Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giró-i-Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language. In *Conference on Computer Vision and Pattern Recognition*, 2021.

Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. Signllm: Sign language production large language models. In *IEEE/CVF International Conference on Computer Vision - Workshops*, 2025a.

Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Yapeng Tian, and Chen Chen. Signdiff: Diffusion model for american sign language production. In *19th IEEE International Conference on Automatic Face and Gesture Recognition, FG*, 2025b.

Sen Fang, Yalin Feng, Hongbin Zhong, Yanxin Zhang, and Dimitris N. Metaxas. Stable signer: Hierarchical sign language generative model. In *Annual Meeting of the Association for Computational Linguistics*, 2026.

Manuel Gil-Martín, María Villa-Monedero, Andrzej Pomirski, Daniel Sáez-Trigueros, and Rubén San Segundo. Sign language motion generation from sign characteristics. *Sensors*, 2023.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Conference on Computer Vision and Pattern Recognition*, 2025.

Eui Jun Hwang, Jung-Ho Kim, and Jong C. Park. Non-autoregressive sign language production with gaussian space. In *British Machine Vision Conference*, 2021.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *Conference and Workshop on Neural Information Processing Systems*, 2023.

Jinlai Liu, Jian Han, Bin Yan, Hui Wu, Fengda Zhu, Xing Wang, Yi Jiang, Bingyue Peng, and Zehuan Yuan. Infinitystar: Unified spacetime autoregressive modeling for visual generation. In *Conference and Workshop on Neural Information Processing Systems*, 2025.

Xiaohan Ma, Rize Jin, and Tae-Sun Chung. Multi-channel spatio-temporal transformer for sign language production. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING*, 2024.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations*, 2024.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition*, 2019.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Conference on Computer Vision and Pattern Recognition*, 2025.

Fan Qi, Yu Duan, Huaiwen Zhang, and Changsheng Xu. Signgen: End-to-end sign language video generation with latent diffusion. In *European Conference on Computer Vision*, 2024.

Guanren Qiao, Sixu Lin, Ronglai Zuo, Zhizheng Wu, Kui Jia, and Guiliang Liu. Signbot: Learning human-to-humanoid sign language interaction. In *International Conference on Robotics and Automation*, 2026.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 2017.

Ben Saunders, Richard Bowden, and Necati Cihan Camgöz. Adversarial training for multi-channel sign language production. In *31st British Machine Vision Conference 2020, BMVC*, 2020a.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *European Conference on Computer Vision*, 2020b.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *Int. J. Comput. Vis.*, 2020.

Stephanie Stoll, Armin Mustafa, and Jean-Yves Guillemaut. There and back again: 3d sign language generation from text using back-translation. In *International Conference on 3D Vision, 3DV*, 2022.

Alexandre Symeonidis-Herzig, JianHe Low, Ozge Mercanoglu Sincan, and Richard Bowden. M3T: discrete multi-modal motion tokens for sign language production. *arXiv preprint*, 2026.

Xinyu Tan, Ningwei Bai, Harry Gardener, Zhengyang Zhong, Luoyu Zhang, Liuhaichen Yang, Zhekai Duan, Monkgogi Galeitsiwe, and Zezhi Tang. Signvla: A gloss-free vision-language-action framework for real-time sign language-guided robotic manipulation. *arXiv*, 2026.

IBM Granite Speech Team. Granite 4.1 speech, 2026. URL https://huggingface.co/ibm-granite/granite-speech-4.1-2b.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Conference and Workshop on Neural Information Processing Systems*, 2024.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Conference and Workshop on Neural Information Processing Systems*, 2017.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, 2021.

Aoxiong Yin, Haoyuan Li, Kai Shen, Siliang Tang, and Yueting Zhuang. T2S-GPT: dynamic vector quantization for autoregressive sign language production from text. In *Annual Meeting of the Association for Computational Linguistics*, 2024a.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, Ziwei Liu, and Lei Yang. WHAC: world-grounded humans and cameras. In *European Conference on Computer Vision*, 2024b.

Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. In *European Conference on Computer Vision*, 2024.

Zhiwei Zheng, Shibo Jin, Lingjie Liu, and Mingmin Zhao. Next-scale autoregressive models for text-to-motion generation. In *Conference on Computer Vision and Pattern Recognition*, 2026.

Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as tokens: A retrieval-enhanced multilingual sign language generator. In *International Conference on Computer Vision*, 2025.

A Appendix

A.1 Multi-Scale Construction and Cross-Scale Conditioning

Temporal pyramid construction. Let $L = L_M$ denote the finest-scale FSQ sequence length. For temporal divisor $d_m \in \{8, 4, 2, 1\}$, the target scale length is

$$L_m = \max\left(1, \left\lceil \frac{L}{d_m} \right\rceil\right). \tag{7}$$

The body, left-hand, and right-hand streams are processed independently using their corresponding FSQ code vectors, while sharing the same target length L_m. Every coarser scale is constructed directly from the finest-scale sequence rather than recursively from the immediately finer scale.

For output position $r = 0, \ldots, L_m - 1$, adaptive average pooling uses the interval

$$a_r = \left\lfloor \frac{rL}{L_m} \right\rfloor, \qquad b_r = \left\lceil \frac{(r+1)L}{L_m} \right\rceil, \tag{8}$$

and computes

$$\bar{\mathbf{q}}_r^{(m)} = \frac{1}{b_r - a_r} \sum_{i=a_r}^{b_r-1} \mathbf{q}_i, \tag{9}$$

where $\mathbf{q}_i$ is the dequantized FSQ code vector at finest-scale position i. Each pooled vector is subsequently assigned to its nearest valid FSQ code. Hence, when L is not divisible by the temporal divisor, no tail tokens are discarded and no zero padding is introduced. For $d_m = 1$, the finest-scale token sequence is copied directly without pooling or requantization.

Cross-scale alignment and gating. Cross-scale context alignment is distinct from the adaptive pooling used to construct supervision targets. For a preceding scale $k < m$, target position t at scale m gathers source position

$$j(t) = \left\lfloor \frac{tL_k}{L_m} \right\rfloor, \tag{10}$$

such that

$$[\mathcal{A}_{k\to m}(E_k)]_t = (E_k)_{j(t)}. \tag{11}$$

The gate logits form a learned matrix $A \in \mathbb{R}^{M\times M}$. For $k < m$,

$$g_{m,k} = \sigma(A_{m,k}), \tag{12}$$

and the cross-scale context is

$$H_m = W_m \left(\frac{\sum_{k<m} g_{m,k}\, \mathcal{A}_{k\to m}(E_k)}{\sum_{k<m} g_{m,k}} \right). \tag{13}$$

Here each $g_{m,k}$ is a single scalar shared across samples, temporal positions, and embedding dimensions, and W_m is a scale-specific bias-free linear projection. The gate logits are initialized to zero, so $g_{m,k} = 0.5$ initially and all available preceding scales contribute equally after normalization.

For the coarsest scale $m = 1$, where no preceding scale is available, a learned seed vector $\mathbf{s}_1$ is repeated over temporal positions and projected as

$$H_1 = W_1 \mathbf{s}_1. \tag{14}$$

A.2 Sign Retrieval

We construct the lexical memory from the isolated-word training split of `akasheroor/American-Sign-Language-Dataset`, resulting in 2,206 dictionary entries. For each lexical item, all available training instances are encoded using the same part-wise FSQ tokenizer as the main model, and the instance with the lowest tokenizer reconstruction MAE is retained as its prototype. Each body, left-hand, and right-hand code sequence is deterministically normalized to 10 FSQ tokens.

Retrieval is implemented as lexical dictionary lookup rather than embedding-based nearest-neighbor search. Given an input sentence, words are normalized to uppercase alphanumeric forms, with a lightweight suffix-based lemma used as fallback. Exact dictionary matches are collected in their original sentence order, duplicate entries are removed, and the first three matches are retained:

$$R(x) = \mathrm{FirstK}\left(\{q_i \mid q_i \in \mathcal{D}\}\right), \qquad K = 3. \tag{15}$$

Thus, $0 \le |R(x)| \le 3$. Each retrieved prototype contributes 10 body, 10 left-hand, and 10 right-hand FSQ codes, for at most 90 raw retrieved code IDs per sentence. These codes are serialized into the textual mT5 conditioning prompt and are used only as auxiliary lexical motion cues; they are not concatenated to the generated motion.

During training, retrieval augmentation is independently dropped for each sample with probability 0.10 to reduce dependence on dictionary coverage. Validation and inference use retrieval without this dropout.

A.3 Self-Conditioning and Context Corruption

SC and GCC modify only preceding-scale contexts during training and do not alter the current-scale supervision. After the teacher-forced forward pass at scale m, we obtain deterministic part-wise predictions by restricted-vocabulary argmax:

$$\hat{z}_\tau^{(m,p)} = \arg\max_{v \in \mathcal{V}_p} \ell_{\tau,v}^{(m,p)}, \tag{16}$$

with vocabulary sizes $|\mathcal{V}_B| = 96$ and $|\mathcal{V}_{LH}| = |\mathcal{V}_{RH}| = 192$. For each sample and completed scale, a single Bernoulli draw with probability p_{SC} determines whether the entire three-part ground-truth scale is replaced by these predicted codes. Thus, the SC decision is shared across all temporal positions and body/hand parts.

GCC is applied after SC. Each valid context token is independently selected with probability p_{GCC} and replaced by a uniformly sampled code from the corresponding part vocabulary. Padding positions are never corrupted.

Both probabilities are scheduled by optimizer update s:

$$\rho(s; p_{\max}, w, r) = \begin{cases} 0, & s < w, \\ p_{\max} \min\left(1, \frac{s-w}{r}\right), & s \ge w. \end{cases} \tag{17}$$

We use

$$p_{\mathrm{SC}}(s) = \rho(s; 0.15, 2000, 6000), \qquad p_{\mathrm{GCC}}(s) = \rho(s; 0.12, 1000, 5000). \tag{18}$$

The same probabilities are used at all scales. Neither SC nor GCC is used during inference, where deterministic argmax decoding is used for benchmark evaluation.

A.4 Training Objectives

For scale m and part p, token cross-entropy is computed over valid (non-padding) positions:

$$\mathcal{L}_{\mathrm{CE}}^{(m,p)} = -\frac{1}{N_{m,p}} \sum_{(i,t) \in \Omega_{m,p}} \log p_\theta\left(z_{i,t}^{(m,p)} \mid x_i, Z_i^{(<m)}, Z_{i,<t}^{(m)}\right). \tag{19}$$

Token CE uses no label smoothing. We use four temporal divisors $(8, 4, 2, 1)$ with raw scale weights

$$(w_1, w_2, w_3, w_4) = (1.25, 1.10, 1.00, 1.25),$$

and body/left-hand/right-hand coefficients $(1.0, 0.625, 0.625)$. The complete training objective is

$$\begin{aligned} \mathcal{L} = \frac{1}{4.60} \sum_{m=1}^{4} w_m \Big[\mathcal{L}_{\mathrm{CE}}^{(m,B)} + 0.625\, \mathcal{L}_{\mathrm{CE}}^{(m,LH)} + 0.625\, \mathcal{L}_{\mathrm{CE}}^{(m,RH)} \Big] \\ + 0.20\, \mathcal{L}_{\mathrm{len}} + 0.03\, \mathcal{L}_{\mathrm{con}}, \end{aligned} \tag{20}$$

where $4.60 = \sum_m w_m$ normalizes the scale-weighted token loss.

Length prediction. Let $L_i \in \{1, \dots, 100\}$ denote the ground-truth finest-scale token length. The model predicts a 100-way distribution $p^{\text{len}}_{i,k}$, with target

$$q_{i,k} = (1 - \epsilon_{\text{len}})\mathbf{1}[k = L_i - 1] + \frac{\epsilon_{\text{len}}}{100}, \qquad \epsilon_{\text{len}} = 0.02. \tag{21}$$

The length loss is

$$\mathcal{L}_{\text{len}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=0}^{99} q_{i,k} \log p^{\text{len}}_{i,k}. \tag{22}$$

At inference, the finest-scale token length is predicted by the maximum-probability length class.

Text–motion contrastive loss. We obtain ℓ_2-normalized text and motion embeddings t_i and m_i, respectively. The text representation is mean-pooled from the language encoder, while the motion representation is constructed from the ground-truth finest-scale tokens. Body, left-hand, and right-hand token embeddings are combined with weights $(0.2, 0.4, 0.4)$ and temporally mean-pooled. Defining

$$s_{ij} = \frac{t_i^\top m_j}{\tau}, \qquad \tau = 0.07, \tag{23}$$

we use the symmetric InfoNCE objective

$$\mathcal{L}_{\text{con}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log \frac{\exp(s_{ii})}{\sum_j \exp(s_{ij})} + \log \frac{\exp(s_{ii})}{\sum_j \exp(s_{ji})}\right]. \tag{24}$$

Other examples in the local minibatch serve as negatives.

Optimization details. Each part tokenizer uses a temporal convolutional encoder–decoder with width and latent dimension 512, two $\times 2$ temporal downsampling/upsampling stages, and residual-stack depth 3, yielding a $4\times$ temporal compression. Each tokenizer is optimized using masked Smooth-L_1 reconstruction loss plus an FSQ commitment loss with weight 0.02. The body levels $\{4, 4, 6\}$ yield 96 indices, while the hand levels $\{4, 6, 8\}$ yield 192 indices per hand. The tokenizers are trained for up to 500 epochs using AdamW with learning rate 2×10^{-4}, gradient clipping at 1.0, and bfloat16 precision. The generator is optimized for at most 100,000 updates using learning rates of 2×10^{-5} for mT5 and 10^{-4} for newly introduced modules, with an effective batch size of 192 on eight NVIDIA A100 GPUs. Checkpoints are selected by generation validation with early stopping after ten evaluations without improvement.

A.5 Robotic Retargeting

Following task-space human-to-robot transfer (Qiao et al., 2026; Chang et al., 2024), the recovered SMPL-X upper-body targets are expressed in a torso-centered coordinate frame and scaled according to the target robot morphology. Let $\mathbf{y}_t$ denote the elbow, wrist, and orientation targets at frame t. The robot configuration is obtained by solving

$$q_t^\star = \arg\min_{q \in \mathcal{Q}} E_{\text{task}}(q; \mathbf{y}_t) + \lambda_{\text{post}} E_{\text{post}}(q) + \lambda_{\text{temp}} \|q - q_{t-1}^\star\|_2^2, \tag{25}$$

where $\mathcal{Q}$ enforces joint limits, E_{task} matches the human task-space targets, and E_{post} resolves kinematic ambiguity by encouraging appropriate elbow configurations and bimanual relations. The preceding-frame solution is used for initialization to promote temporal continuity.

Hand articulation is retargeted independently by matching palm-normalized MANO finger geometry to the robot hand under joint-limit and temporal-continuity constraints. The resulting body and hand references are temporally regularized and tracked by a platform-specific low-level controller. Motion channels unsupported by the target embodiment are omitted during retargeting.

A.6 Physical Embodiment Evaluation

We quantitatively evaluate the embodiment stage on all 30 available retargeted sequences, comprising 15 ground-truth (GT) motions and the corresponding 15 motions generated by RoBoSTAR, for a total of 11,890 scored retargeting frames.

Table 5: **Quantitative physical-embodiment evaluation.** Retargeting measures morphology-transfer error, while execution measures closed-loop tracking in physics simulation. Results are mean $\pm$ standard deviation across sequences.

Stage	Motion	Elbow (cm)	Wrist (cm)	Wrist ori. (°)	Joint MAE (°)
Retargeting	GT	2.0 ± 0.4	4.6 ± 1.4	22.9 ± 10.9	–
	Ours	2.0 ± 0.5	5.6 ± 1.9	31.3 ± 17.7	–
Execution (sim.)	GT	1.6 ± 0.5	3.6 ± 0.9	10.3 ± 3.2	2.1 ± 0.7
	Ours	0.9 ± 0.3	2.1 ± 0.6	5.9 ± 1.6	1.3 ± 0.4

Metrics. Retargeting fidelity is evaluated in the same robot-centered task space used by the retargeting objective, after applying the corresponding torso-frame transformation and morphology scaling. For task-space joint j at frame t, we compute

$$e_{j,t}^{\mathrm{pos}} = \left\| p_j(q_t^\star) - y_{j,t} \right\|_2 , \tag{26}$$

where $y_{j,t}$ is the morphology-mapped target and $p_j(q_t^\star)$ is the position obtained by forward kinematics from the retargeted robot configuration. We report elbow and wrist position error, together with the geodesic orientation error of the wrist.

We further evaluate controller-level tracking using closed-loop MuJoCo simulation with the same torque-PD controller used by the execution pipeline. Tracking error is measured between the commanded retargeted reference and the simulated configuration using task-space errors and joint-angle MAE. Since encoder-level logs are unavailable for the physical demonstrations, these values quantify simulated closed-loop tracking rather than measured hardware tracking.

Kinematic feasibility. The commanded retargeted configurations remain within the enforced joint ranges. Because the retargeter explicitly clips configurations to these ranges, the zero commanded-violation rate is expected; more informatively, 13.8% of arm-joint entries lie on a joint bound. During closed-loop simulation, 1.67% of executed joint entries transiently exceed the configured range, with a maximum excess of $9.5°$, primarily at wrist-roll and finger joints under self-contact.

Discussion. The generated motions exhibit elbow retargeting error identical to GT motion and only a modest increase in wrist position error (5.6 cm versus 4.6 cm), indicating that the generated trajectories remain compatible with the target robot morphology. Wrist orientation is the dominant residual, suggesting that orientation-aware retargeting is a remaining source of embodiment error. Closed-loop simulation further shows that the resulting references can be tracked with centimeter-level end-effector error and low joint-angle error.

The archived controller rollouts use a fixed 20-fps source rate, whereas the original sequences have heterogeneous native frame rates. This does not affect the frame-indexed retargeting or commanded joint-limit metrics, but may make the simulated tracking errors optimistic. We therefore use the simulation results as evidence of controller-level feasibility rather than measured hardware fidelity.

A.7 Evaluation Details

Geometric metrics. Let P_i and G_j denote the predicted and reference 3D joint positions at frames i and j, respectively. For $p \in \{B, LH, RH\}$, let $K_B = 12$ and $K_{LH} = K_{RH} = 21$. For DTW-JPE, body predictions are translation-aligned at the SMPL-X root, while each hand is aligned at its wrist. Denoting the aligned prediction by $\widetilde{P}^p_{i,k;j}$, the pairwise cost is

$$c^p_{\mathrm{JPE}}(i,j) = \frac{1}{K_p} \sum_{k=1}^{K_p} \left\| \widetilde{P}^p_{i,k;j} - G^p_{j,k} \right\|_2 . \tag{27}$$

For DTW-PA-JPE, we instead apply a per-frame similarity Procrustes transform

$$\mathcal{A}_{ij}(z) = szR^\top + t, \tag{28}$$

which includes translation, rotation, and uniform scale, and define

$$c^p_{\mathrm{PA}}(i,j) = \frac{1}{K_p}\sum_{k=1}^{K_p}\left\|\mathcal{A}_{ij}(P^p_{i,k}) - G^p_{j,k}\right\|_2. \tag{29}$$

For the body, the similarity transform is fitted using the SMPL-X joints and evaluated on the 12 upper-body joints; for each hand, it is fitted and evaluated on the corresponding 21 points.

For either pairwise cost c, standard monotonic DTW determines the minimum-cost alignment path $\pi^\star$. The reported score is

$$E^p_{\mathrm{DTW}} = \sum_{(i,j)\in\pi^\star} c^p(i,j), \qquad E_{\mathrm{hand}} = \frac{1}{2}\left(E_{\mathrm{left}} + E_{\mathrm{right}}\right). \tag{30}$$

All geometric distances are measured in meters. We report the mean per-sequence cumulative DTW cost over the test set.

Motion dynamics. Dynamics are computed on the native-length z-normalized 133D motion representation. For feature dimension d,

$$\widetilde{x}_{i,t,d} = \frac{x_{i,t,d} - \mu_d}{\max(\sigma_d, 10^{-8})}. \tag{31}$$

We apply a centered five-frame moving average to obtain $\bar{\mathbf{x}}_{i,t}$ and define the transition magnitude

$$r_{i,t} = \sqrt{\frac{1}{D}\left\|\bar{\mathbf{x}}_{i,t+1} - \bar{\mathbf{x}}_{i,t}\right\|_2^2}, \qquad D = 133. \tag{32}$$

The motion threshold is calibrated from ground-truth sequences as

$$\tau = Q_{0.20}\left(\left\{r^{\mathrm{GT}}_{i,t}\right\}_{i,t}\right). \tag{33}$$

Dynamic Degree is the fraction of transitions exceeding this threshold, macro-averaged across sequences:

$$\mathrm{DynamicDegree} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i - 1}\sum_{t=1}^{T_i-1}\mathbf{1}[r_{i,t} > \tau]. \tag{34}$$

Because τ is the 20th percentile of GT transition magnitudes, the reference value is approximately 0.8.

For sequence i, define its mean motion magnitude as

$$M_i = \frac{1}{T_i - 1}\sum_{t=1}^{T_i-1} r_{i,t}. \tag{35}$$

Magnitude Ratio is the ratio between the dataset-level mean predicted and GT magnitudes:

$$\mathrm{MagnitudeRatio} = \frac{\frac{1}{N}\sum_i M^{\mathrm{pred}}_i}{\frac{1}{N}\sum_i M^{\mathrm{GT}}_i}. \tag{36}$$

Its reference value is therefore 1.

For Freeze-tail, let

$$K^{\mathrm{tail}}_i = \max\left\{k : r_{i,T_i-j} \le \tau,\ j = 1,\dots,k\right\} \tag{37}$$

denote the number of consecutive low-motion transitions at the end of sequence i. We define

$$\mathrm{FreezeTail} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left[\frac{K^{\mathrm{tail}}_i}{T_i - 1} \ge 0.25\right]. \tag{38}$$

Thus, Freeze-tail is the fraction of sequences whose terminal static run occupies at least 25% of all transitions. Dynamic Degree, Magnitude Ratio, and Freeze-tail are dimensionless.

Table 6: Effect of the language-model backbone on motion generation.

Backbone	DTW-PA-JPE ↓		DTW-JPE ↓	
	Body	Hand	Body	Hand
mBART	4.38	1.88	4.24	5.95
mT5-Base	4.21	1.93	4.07	6.05
mT5-Large	**3.96**	**1.52**	**3.87**	**5.34**

Table 7: Effect of sign retrieval on motion generation.

Configuration	DTW-PA-JPE ↓		DTW-JPE ↓	
	Body	Hand	Body	Hand
w/o Retrieval	4.68	1.91	4.58	6.10
w/ Retrieval	**3.96**	**1.52**	**3.87**	**5.34**

Back-translation evaluator. For semantic evaluation, we train a separate Hybrid FSQ-T5 back-translator on the official How2Sign splits, comprising 28,691 training, 1,578 validation, and 2,167 test sequences. The evaluator combines frozen body and bilateral-hand FSQ codes with continuous four-frame motion statistics through gated fusion and conditions a Flan-T5-base sequence-to-sequence model. The continuous features capture the mean, endpoint displacement, and standard deviation within each four-frame chunk. Auxiliary content-word CTC, semantic-distillation, and masked-FSQ objectives are used during training.

The checkpoint is selected by validation corpus BLEU-4 and subsequently frozen for all comparisons. On the held-out How2Sign test split, back-translation from ground-truth motion achieves BLEU-1, -2, -3, and -4 scores of 22.45, 8.59, 3.95, and 1.96, respectively. For back-translation only, generated motions are linearly resampled to the corresponding reference length before being passed to the evaluator; geometry and motion-dynamics metrics are computed at their native generated lengths. Since the back-translator itself is imperfect, these scores are used as a comparative measure of semantic consistency rather than an absolute measure of translation accuracy.

A.8 Additional Ablations

Language-model backbone. Table 6 compares language-model backbones while keeping the motion representation, generator, and training protocol unchanged. mT5-Large achieves the lowest errors across all four geometric metrics, with particularly pronounced improvements in hand motion. Compared with mT5-Base, it reduces DTW-PA-JPE by 5.9% and 21.2% for the body and hands, respectively, and DTW-JPE by 4.9% and 11.7%.

Sign retrieval. Inspired by SOKE (Zuo et al., 2025), we augment the text condition with retrieved word-level sign tokens. As shown in Table 7, retrieval improves all four geometric metrics, with a particularly clear gain in hand motion. Retrieved signs serve only as auxiliary lexical cues; sentence-level motion remains generated autoregressively rather than assembled from isolated sign clips.